Multi-Paced Dictionary Learning for Cross-Domain Retrieval and Recognition

Dan Xu, Jingkuan Song, Xavier Alameda-Pineda, Elisa Ricci, Nicu Sebe

Multi-Paced Dictionary Learning for Cross-Domain Retrieval and Recognition

Dan Xu[1], Jingkuan Song[1], Xavier Alameda-Pineda[1], Elisa Ricci[2], Nicu Sebe[1]
[1]University of Trento, Italy
[2]Fondazione Bruno Kessler and University of Perugia, Italy

***Abstract*—Several applications benefit from learning coupled representations able to describe data from multiple sources. For instance, cross-domain dictionary learning methods demonstrated to be particularly effective. In this paper we introduce *Multi-Paced Dictionary Learning (MPDL)* and propose an instantiation of it under the framework of cross-domain dictionary learning. MPDL is inspired by previous works on self-paced learning, a framework able to enhance the accuracy of conventional learning models by presenting the training data in a meaningful order, *i.e.* easy samples are provided first. However, most of existing self-paced learning methods only consider a single modality, while MPDL is specifically designed to assess the learning pace when data from multiple sources are available. We present the model and propose an efficient algorithm to learn the dictionaries and codes. The approach is validated via experiments on two different tasks, namely cross-media retrieval and sketch-to-photo face recognition, using publicly available datasets.**

I. Introduction

Many problems in computer vision, natural language processing and multimedia analysis require solving cross-modal knowledge association tasks. Dictionary learning approaches [1], [2], [3] have proven to be successful for learning coupled representations using data from multiple domains. One recurrent issue with latent variable models, and with cross-modal dictionary learning in particular, is that the learning algorithms usually find sub-optimal solutions corresponding to *bad* local minima of the objective function.

Recently, self-paced learning (SPL) [4] has raised as the framework insignia within which many classical learning techniques have shown to improve the generalization performance. The idea of self-paced learning is inspired from the way a good teacher instills knowledge to students: prioritizing easy training samples over the more difficult ones. Since the learning priority is understood from the data themselves, we are left with the challenging task of designing a meaningful strategy to assess the difficulty of the samples. Previous works have addressed this issue in different ways. The most common strategy is to measure the easiness of a sample by its loss [4]. Alternatively, Jiang *et al.* [5] defined the sample order taking into account the dissimilarity with respect to what has already been learned. A hybrid strategy, where prior knowledge (*i.e.* a curriculum [6]) is integrated into the computation of the easiness measure, was presented in [7].

In this paper we address the problem of inferring the sample order using data coming from multiple sources. Specifically, we introduce *Multi-Paced Dictionary Learning* and propose an instantiation of it under the framework of cross-domain dictionary learning. MPDL is related to both curriculum learning and SPL. As explained above, while in curriculum learning the ordering is set by the "teacher" in advance, in SPL the ordering is data-driven, *i.e.* inferred from the "student's" understanding. The rationale behind MPDL is that, within this teacher-student analogy, most successful students learn from multiple sources. In other words the learning pace is jointly determined by the knowledge gathered from the teacher, textbook, electronic resources, etc. Importantly, the learning order must be decided, not only upon one particular explanation of a concept, but also taking into account how much two interpretations of the same concept from different sources match. In the framework of dictionary learning, this corresponds to evaluating not only the representation power of each domain's dictionary, but also the coherence between the domain-specific codes. Thus, the learning rhythm is determined by multiple sources, giving rise to MPDL.

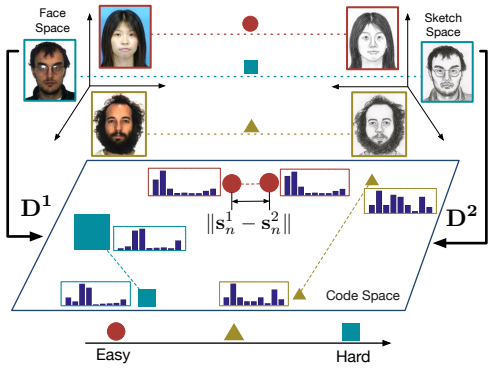


Fig. 1. Application of the multi-paced dictionary learning method to sketch-to-photo face recognition. Faces and sketches are represented as sparse codes by means of the (learned) dictionaries $\mathbf{D}^1$ and $\mathbf{D}^2$. The shape size represents the reconstruction error of the corresponding codes $\mathbf{s}_n^1, \mathbf{s}_n^2$. Intuitively, our method first selects the sample associated with the red circle, then with the yellow triangle and finally with the blue square. Details are given in main text.

Figure 1 illustrates the idea behind the proposed method. MPDL first trains with those samples that have a good trade-off between reconstruction error and code matching. Intuitively, our method first selects the sample associated with the red circle, then with the yellow triangle and finally with the blue square. Even if the yellow codes show low reconstruction error, they do not match. Contrarily, the red codes match very well while keeping a fairly low reconstruction error. Finally, the samples associated with blue codes are selected

at the end, as they correspond to higher reconstruction errors. To assess the appropriateness of our intuition, we derive a novel coupled dictionary learning framework and propose an efficient algorithm to learn the dictionaries and codes and simultaneously infer the optimal sample order. Finally, we evaluate the effectiveness of our framework by conducting an extensive experimental evaluation on publicly available benchmarks for two applications, namely cross-media retrieval and sketch-to-photo face recognition.

II. Related Work

In this section we describe related work on (i) cross-domain dictionary learning and (ii) self-paced and curriculum learning, since multi-paced dictionary learning lies in the cross-road of these two research directions.

In the last few years, several dictionary learning algorithms have been proposed [8], [9], [10], pushing the state-of-the-art in many applications. When data from different domains are available, traditional dictionary learning approaches can be extended to benefit from the information encoded in multiple sources. In [1], corresponding samples from different domains were used to learn per-source dictionaries in a coupled fashion. Wang *et al.* [2] introduced a semi-coupled dictionary learning scheme for cross-domain image sparse representation. Huang and Wang [3] proposed a framework to learn a pair of dictionaries along with the corresponding feature representations. Finally, a multi-task dictionary learning framework was proposed in [11] in the context of the automatic analysis of paintings to learn multiple style-specific dictionary representations.

Coupled dictionary learning strategies have shown to be especially effective in transfer learning tasks, *i.e.* when knowledge obtained from a given dataset (source) is used to facilitate the learning on another dataset (target). For instance, Ding *et al.* [12] presented a framework to handle the *missing modality* problem (target domain data are not available). Similarly, Ni *et al.* [13] used dictionaries to learn a set of subspaces minimizing the distance between the data distributions of source and target domains.

Curriculum [6] and self-paced [4] learning develop from the idea that models must be learned in an incremental fashion, using the easy samples before the difficult ones. However, while curriculum learning requires the *a priori* identification of easy and hard samples in a given training dataset, in self-paced the learning order is automatically determined from the data themselves. Due to its generality, curriculum and self-paced learning have been considered in a broad spectrum of latent variable models, including matrix factorization [14], domain adaptation [15], dictionary learning [16] and clustering [17]. Among all previous works, the two studies most related to this paper are [16] and [17]. In [16] a SPL strategy is proposed in the context of dictionary learning. Opposite to [16], in this paper we analyze the challenging problem of combining data from multiple domains and develop a novel dictionary learning framework for coupling multiple modalities and selecting the sample order accordingly. In [17] the traditional SPL scheme is embedded in an algorithm which clusters multi-view data. As a consequence, the order of datapoints is determined considering each view in isolation, while in our MPDL the samples are sorted by jointly looking at multiple domains.

III. The Model of MPDL

In this section we present our MPDL approach. We first introduce the problem of dictionary learning when samples are gathered from multiple data sources. Then, we describe the proposed MPDL method along with the algorithm we devised to solve the associated optimization problem.

A. Cross-Domain Dictionary Learning

We assume the existence of N samples observed in I different modalities. We denote the feature vector of the n^{th} sample in the i^{th} modality as $\mathbf{x}_n^i \in \mathbb{R}^{M_i}$, where M_i is the dimension of the i^{th} feature space. We choose to instantiate MPDL within the framework of sparse coding and dictionary learning [8], [9], [10]. More precisely, we assume the existence of I dictionaries – one per modality – that represent the data as $\mathbf{x}_n^i \approx \mathbf{D}^i \mathbf{s}_n^i$, where $\mathbf{s}_n^i \in \mathbb{R}^K$ is the code of the n^{th} sample in the i^{th} modality and $\mathbf{D}^i \in \mathbb{R}^{M_i \times K}$ is the dictionary of the i^{th} modality containing K words. Compactly, we write:

$$\mathbf{X}^i \approx \mathbf{D}^i \mathbf{S}^i, \tag{1}$$

where $\mathbf{X}^i = [\mathbf{x}_1^i, \dots, \mathbf{x}_N^i]$ and $\mathbf{S}^i = [\mathbf{s}_1^i, \dots, \mathbf{s}_N^i]$. Dictionary learning boils down to a loss minimization problem with normalization constraints and sparse regularization:

$$\begin{aligned} \min_{\{\mathbf{S}^i, \mathbf{D}^i\}_{i=1}^I} & \sum_{i=1}^{I} \left(\left\| \mathbf{X}^i - \mathbf{D}^i \mathbf{S}^i \right\|_2^2 + \alpha \left\| \mathbf{S}^i \right\|_1 \right) \\ & \text{s.t. } \|\mathbf{d}_k^i\| \leq 1, \end{aligned} \tag{2}$$

where $\mathbf{d}_k^i$ is the k^{th} word of the i^{th} dictionary, *i.e.* , $\mathbf{D}^i = [\mathbf{d}_1^i, \dots, \mathbf{d}_K^i]$. The constraint allows to skip scale ambiguities due to the matrix product $\mathbf{D}^i \mathbf{S}^i$.

Prior relational knowledge among the samples is very useful when learning new representations [18], [19]. To embed this knowledge into dictionary learning, we propose to use the original set of features $\mathbf{X}^i$ to create an undirected proximity graph. In practice, the weights of this graph w_{nm}^i can be computed, for instance, with the Gaussian kernel. The weights are then used to create the Laplacian matrix $\mathbf{L}^i$ with $l_{mn}^i = \sum_{p=1}^N w_{np}^i - w_{nm}^i$. A regularization term is then added to the dictionary problem (2) leading to:

$$\begin{aligned} \min_{\{\mathbf{S}^i, \mathbf{D}^i\}_{i=1}^I} & \sum_{i=1}^{I} \left(\left\| \mathbf{X}^i - \mathbf{D}^i \mathbf{S}^i \right\|_2^2 + \alpha \left\| \mathbf{S}^i \right\|_1 + \beta \text{Tr} \left(\mathbf{S}^i \mathbf{L}^i {\mathbf{S}^i}^\top \right) \right) \\ & \text{s.t. } \|\mathbf{d}_k^i\| \leq 1. \end{aligned} \tag{3}$$

The problem (3) learns a set of domain-specific dictionaries along with the associated code independently for each domain. Previous works [2], [1], [3] have shown that a coupled learning framework can be beneficial in several applications. Our MPDL approach develops from this intuition. However, opposite to previous works, in MPDL modality-specific dictionaries and codes are learned jointly by presenting the training data

in a meaningful order, *i.e.* easy samples are provided first. Our experiments demonstrate that the proposed strategy is superior to previous multi-domain dictionary learning methods [2], [3].

B. The Formulation of MPDL

As in curriculum and self-paced learning, MPDL develops from the idea that models must be learned in an incremental fashion, using the easy samples before the difficult ones. To model the easiness of learning the n^{th} sample in MPDL a binary variable $v_n \in \{0,1\}$ is introduced that indicates whether or not the n^{th} sample is currently selected for training. Consequently only the samples selected for training are taken into account and the i^{th} dictionary loss becomes $\left\|(\mathbf{X}^i - \mathbf{D}^i\mathbf{S}^i)\mathbf{V}\right\|_2^2$, where $\mathbf{V}$ is a N-by-N diagonal matrix with entries $v_1, \ldots, v_N$. As it is traditionally done, the method iteratively increases the amount of samples selected for learning by adding a regularization term with a penalty evolving over time [4], [5], [7]. In the present study, we assume that the data is split into C groups (either provided or learned from the data) and define a group-specific indicator vector $\mathbf{p}_c \in \mathbb{R}^N$, where $p_{c,n} = 1$ if and only if sample n belongs to group c, and $p_{c,n} = 0$ otherwise. We devise a penalty over $\mathbf{V}$ that is normalized over the groups' size, denoted by B_c:

$$\sum_{c=1}^{C} \frac{1}{B_c} \|\mathbf{V}\mathbf{p}_c\|_1. \tag{4}$$

Importantly, the penalty induced by this term has to evolve over time so to allow more and more samples to be part of the training set. At the same time, this term enforces learning from different groups and therefore it is closely related to SPL with diversity [20]. Similarly to [20], the idea is to learn not only from easy samples as in standard SPL [4] but also from samples that are dissimilar from what has already been learned. However, with respect to [20], the proposed regularizer has two prominent advantages:(i) we avoid using group norms that significantly increase the complexity of the optimization solvers and (ii) we introduce the normalization factor B_c that softens the bias induced by dissimilar group cardinalities. In all:

$$\begin{aligned}\min_{\mathbf{V},\{\mathbf{S}^i,\mathbf{D}^i\}_{i=1}^I} \sum_{i=1}^{I} \Big(\left\|(\mathbf{X}^i - \mathbf{D}^i\mathbf{S}^i)\mathbf{V}\right\|_2^2 + \alpha \left\|\mathbf{S}^i\right\|_1 \\ + \beta \mathrm{Tr}\left(\mathbf{S}^i\mathbf{L}^i{\mathbf{S}^i}^\top\right)\Big) - \mu \sum_{c=1}^{C} \frac{1}{B_c} \|\mathbf{V}\mathbf{p}_c\|_1 \\ \text{s.t. } \|\mathbf{d}_k^i\| \leq 1 \text{ and } v_n \in \{0,1\}.\end{aligned} \tag{5}$$

As delineated in the introduction, MPDL aims to determine the learning pace from multiple sources, that is, from the different I domains. Importantly, MPDL assesses the learning easiness of the n^{th} sample, not only from the representation precision in each of the I domains (as is already done in (5)), but also from the correspondence between cross-domain representations. This is the main conceptual contribution of the present research study. Moreover, within the framework of dictionary learning this is naturally implemented as a penalty over the codes' dissimilarity. Importantly, this dissimilarity is also weighted by the pacing variables $\mathbf{V}$. The regularization term is defined as:

$$\sum_{i<j} \| \left(\mathbf{S}^i - \mathbf{S}^j\right) \mathbf{V}\|_2^2. \tag{6}$$

Finally, the optimization problem for MPDL writes:

$$\begin{aligned}\min_{\mathbf{V},\{\mathbf{S}^i,\mathbf{D}^i\}_i} \sum_i \Big(\left\|(\mathbf{X}^i - \mathbf{D}^i\mathbf{S}^i)\mathbf{V}\right\|_2^2 + \alpha \left\|\mathbf{S}^i\right\|_1 \\ + \beta \mathrm{Tr}\left(\mathbf{S}^i\mathbf{L}^i{\mathbf{S}^i}^\top\right)\Big) - \mu \sum_c \frac{\|\mathbf{V}\mathbf{p}_c\|_1}{B_c} + \gamma \sum_{i<j} \| \left(\mathbf{S}^i - \mathbf{S}^j\right) \mathbf{V}\|_2^2 \\ \text{s.t. } \|\mathbf{d}_k^i\| \leq 1 \text{ and } v_n \in \{0,1\},\end{aligned} \tag{7}$$

where the regularization parameters $\alpha, \beta, \mu, \gamma$ balance the effect of sparseness, the prior knowledge, the general pacing regime and the plurality of the learning pace, respectively.

Algorithm 1 Multi-Paced Dictionary Learning

Input: Initialize $\mathbf{D}^i$, $\mathbf{S}^i$ using single modality dictionary learning [1], μ to use 10% of the training set and the parameters α, β, γ;

```
repeat
    for i = [1, ..., I] do
        Update S^i according to (10);
        Update D^i by solving (8);
    end for
    Update the v_n's following (14);
    Increase μ by λ to enlarge the training set;
until All the training data points are selected.
```

Output: $\mathbf{D}^i$, $\mathbf{S}^i$, $\mathbf{V}$;

C. The Optimization of MPDL

The optimization problem for MPDL is not jointly convex. We propose an alternating optimization approach to solve (7) in three steps: dictionaries, codes and pacing variables.

Solve for $\mathbf{D}^i$: With fixed codes $\mathbf{S}^i$ and selection matrix $\mathbf{V}$, the optimization problem for $\mathbf{D}^i$ writes:

$$\min_{\mathbf{D}^i} \left\|(\mathbf{X}^i - \mathbf{D}^i\mathbf{S}^i)\mathbf{V}\right\|_2^2 \quad \text{s.t. } \left\|\mathbf{d}_k^i\right\| \leq 1. \tag{8}$$

This problem is a Quadratically Constrained Quadratic Program (QCQP) that can be solved using gradient descent with iterative projection [21] or Lagrangian duality, *e.g.* [10].

Solve for $\mathbf{S}^i$: With fixed dictionaries $\mathbf{D}^i$ and selection matrix $\mathbf{V}$, the optimization function for the codes writes:

$$\begin{aligned}f(\mathbf{S}^i) = \sum_{i=1}^{I} \Big(\left\|\left(\mathbf{X}^i - \mathbf{D}^i\mathbf{S}^i\right)\mathbf{V}\right\|_2^2 + \alpha \left\|\mathbf{S}^i\right\|_1 \\ + \beta \mathrm{Tr}\left(\mathbf{S}^i\mathbf{L}^i{\mathbf{S}^i}^\top\right)\Big) + \gamma \sum_{i<j} \left\|\left(\mathbf{S}^i - \mathbf{S}^j\right)\mathbf{V}\right\|_2^2\end{aligned} \tag{9}$$

According to FISTA [22], (9) can be viewed as a proximal regularization problem, iteratively solved using the following recursion (over r):

$$\mathbf{S}_r^i = \underset{\mathbf{S}^i}{\operatorname{argmin}} \left\{ \frac{\left\| \mathbf{S}^i - \left(\mathbf{S}_{r-1}^i - t_r \nabla f\left(\mathbf{S}_{r-1}^i \right) \right) \right\|_2^2}{2t_r} + \alpha \left\| \mathbf{S}^i \right\|_1 \right\}, \tag{10}$$

where[1]:

$$\begin{aligned} \nabla f(\mathbf{S}^i) = 2\mathbf{D}^{i\top}(\mathbf{D}^i\mathbf{S}^i - \mathbf{X})\mathbf{V} + 2\beta\mathbf{S}^i\mathbf{L}^i \\ + 2\gamma \sum_{j \neq i} (\mathbf{S}^i - \mathbf{S}^j)\mathbf{V} \end{aligned} \tag{11}$$

is the gradient of $f(\mathbf{S}^i)$ and $t_r > 0$ is a suitable step size. Since the ℓ_1 norm is separable, the computation of each $\mathbf{s}_{n,r}^i$ reduces to solving a one-dimensional minimization problem for each of its components, which by simple calculus produces $\mathbf{s}_{n,r}^i = \tau_{\alpha t_r}\left(\mathbf{s}_{n,r-1}^i - t_r \nabla f\left(\mathbf{s}_{n,r-1}^i \right) \right)$, where τ_λ is the shrinkage operator defined by: $\tau_\lambda(\mathbf{x})_m = (|x_m| - \lambda)_+ \operatorname{sgn}(x_m)$ and m is the dimensionality of $\mathbf{x}$.

Solve for $\mathbf{V}$: By fixing the dictionaries $\mathbf{D}^i$ and the codes $\mathbf{S}^i$, we can update $\mathbf{V}$ solving:

$$\begin{aligned} \min_{\mathbf{V}} \sum_{i=1}^{I} \left\| \left(\mathbf{X}^i - \mathbf{D}^i\mathbf{S}^i \right) \mathbf{V} \right\|_2^2 + \gamma \sum_{i<j} \left\| \left(\mathbf{S}^i - \mathbf{S}^j \right) \mathbf{V} \right\|_2^2 \\ - \mu \sum_c \frac{\|\mathbf{V}\mathbf{p}_c\|_1}{B_c}, \quad \text{s.t. } v_n \in \{0,1\}. \end{aligned} \tag{12}$$

Because $\mathbf{V}$ is a diagonal matrix with binary entries, we can rewrite the previous optimization problem as N independent optimization problems of the following form:

$$\min_{v_n} v_n \left(e_n + \gamma f_n - \mu g_n \right) \quad \text{s.t. } v_n \in \{0,1\}, \tag{13}$$

with $e_n = \sum_{i=1}^{I} \|\mathbf{x}_n^i - \mathbf{D}^i\mathbf{s}_n^i\|_2^2$, $f_n = \sum_{i<j} \|\mathbf{s}_n^i - \mathbf{s}_n^j\|_2^2$, $g_n = \frac{1}{B_{c_n}}$, being c_n the class of the n^{th} sample. The solution of the previous optimization problem is trivial and given by:

$$v_n^* = \begin{cases} 1 & \text{if } e_n + \gamma f_n < \mu g_n, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Importantly, this solution should be understood with the following intuition. If the error of a training point is smaller than a threshold μg_n, the training point will be selected for training. To make sure that the training samples are equally selected independently of the clusters, we impose a higher threshold (determined by g_n) for large classes/clusters. The summary of the learning procedure is shown in Algorithm 1.

IV. Experiments

In this section, we extensively evaluate the proposed approach when applied to two tasks, namely cross-media retrieval and sketch-to-photo face recognition.

Fig. 2. Two examples of image query text (the first row) and text query image (the second row). Top four results are retrieved by our method on the Wikipedia dataset.

TABLE I
Mean average precision of all benchmarked methods on different retrieval tasks using the Wikipedia dataset.

Algorithm	Retrieval Task			
	$\mathbb{Q}_{I\to I}$	$\mathbb{Q}_{T\to T}$	$\mathbb{Q}_{I\to T}$	$\mathbb{Q}_{T\to I}$
Random	0.118	0.118	0.118	0.118
CFA	0.157	0.488	0.246	0.195
CCA	0.161	0.495	0.249	0.196
PLS	0.158	0.490	0.240	0.163
LCMH	0.146	0.359	0.133	0.117
CMLSSH	0.148	0.318	0.138	0.140
CVH	0.147	0.153	0.126	0.122
MSAE	0.162	0.462	0.182	0.179
MPDL ($\gamma = 0$)	0.171	0.494	0.237	0.185
MPDL ($\mu \to \infty$)	0.181	0.498	0.269	0.191
MPDL	**0.186**	**0.512**	**0.274**	**0.198**

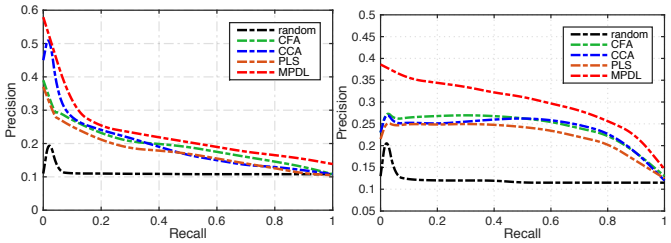
A. Cross-media retrieval

Cross-media retrieval aims at robustly finding query results in various media by submitting queries in one of the media. We use the proposed method to retrieve text and images from the widely used "Wikipedia" dataset [23].

Setup. The Wikipedia dataset consists of 2866 image-text pairs generated from featured articles of Wikipedia. These articles are continuously updated and selected by the Wikipedia editors since 2009. While the text is a semantic description of people, places or events, the images are a visual representation of them: both modalities are complementary. Each text-image pair sample is labeled with one out of 10 possible semantic classes. The dataset is randomly split in 2173 training samples and 693 test samples. Each image is then represented by a 128-dimensional codeword obtained using the bag-of-words framework over SIFT features. The text descriptor is computed by a latent Dirichlet allocation (LDA) model with 10 topics, as in [23]. Following previous works on dictionary learning for retrieval tasks, after training our MPDL, we used the learned dictionaries to compute the codes associated to test samples. These codes are then employed for cross-media retrieval applying a nearest neighbor scheme.

Baselines. We compared the proposed approach with several state-of-the-art methods: Random retrieval, Cross-modal Factor Analysis (CFA) [24], Canonical Correlation Analysis (CCA) [23], Partial Least Squares (PLS) [25], Linear Cross-Modal Hashing (LCMH) [26], Cross-modality Metric Learning with Similarity-Sensitive Hashing (CMLSSH) [27], Cross-View Hashing (CVH) [28] and Multi-modal Stacked Auto-Encoders (MSAE) [29]. Importantly, our method learns the

[1]We used $\mathbf{V}^2 = \mathbf{V}$.



(a) $\mathbb{Q}_{T\to I}$ (b) $\mathbb{Q}_{I\to T}$

Fig. 3. Precision-recall curves of different methods for the cross-modal retrieval tasks.

dictionary in an unsupervised manner, where only pair-wise information is utilized. Semantic information is also important for cross media retrieval, but constructing high-level semantic description usually requires label information [23]. Therefore, we only consider and compare with unsupervised methods. The chosen baselines can be grouped by the way they achieve cross-modal retrieval: CFA and CCA learn modality-specific subspaces; LCMH, CVH, CMLSSH and PLS find a projection matrix by means of spectral hashing schemes or correlation metrics; and MSAE uses two parallel multi-layer stacked auto-encoder networks to learn cross-modal correlations. To guarantee a fair comparison all methods use the same visual features (except MSAE that uses the raw images). The dictionary size was set to $K^1 = K^2 = 7$.

Results. We report mean average precision (MAP) performance of all benchmarked methods on the four possible single- and cross-media retrieval tasks in Table I[2]. The regularization parameters α, β and γ of our method were set by cross-validation to 1, 0.1 and 1, respectively. In addition to the proposed MPDL method with the cross-validated parameters, we also show two special cases, namely $\gamma = 0$, *i.e.* no cross-modal coupling, and $\mu \to \infty$, *i.e.* disabling the self-pacing feature of the method.

We remark that MPDL systematically obtains the best performance among all methods. Specifically, for single-media retrieval, our method obtains MAP scores of 0.185 and 0.512 for the *text-to-text* query task ($\mathbb{Q}_{T\to T}$) and the *image-to-image* query task ($\mathbb{Q}_{I\to I}$), respectively. $\mathbb{Q}_{I\to I}$ shows a significant gap of 2.4 points over MSAE, the best state-of-the-art method. This is probably because MSAE fails to learn joint representations from low-resolution images. For cross-modal retrieval, our method achieves 0.274 for $\mathbb{Q}_{I\to T}$, which is 2.5 points above the best (CCA), and 0.198 for $\mathbb{Q}_{T\to I}$, outperforming all methods. The results show the beneficial effect of assessing the learning pace from cross-modal correspondences, that is when compared to MPDL ($\gamma = 0$). A similar conclusion can be drawn when the learning pace is disabled.

In addition to the MAP, we plot precision-recall (PR) curves for the two cross-modal tasks in Figure 3. The PR results are consistent with Table I, and our method outperforms previous approaches. Finally, we also show some qualitative results in Figure 2, where the images on the right correspond to the first five images selected from the text query on the left.

[2]The numbers corresponding to the performance of the other approaches are partially taken from [29].

Fig. 4. Example of face-sketch pairs in the CUFS dataset.

TABLE II
AVERAGE RECOGNITION RATE FOR ALL BENCHMARKED METHODS ON SKETCH-TO-PHOTO FACE RECOGNITION.

Algorithm	Average Recognition Rate
STM	81.0%
PLS	93.6%
Bil	94.2%
CCA	94.6%
SCDL	95.2%
JDL	95.4%
CDL	97.4%
MPDL($\gamma = 0$)	96.5%
MPDL ($\mu \to \infty$)	97.3%
MPDL	**98.4**%

B. *Sketch-to-photo face recognition*

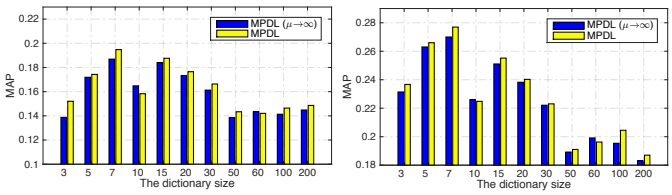
The sketch has become an important clue in image query systems, specially if a picture of the face can be retrieved from the sketch. We therefore applied the proposed method to the sketch-to-photo recognition task. Experiments were conducted on the CUHK Face Sketch Database (CUFS) [30].

Setup. CUFS contains pair-wise samples of sketch and photo faces collected from 188 CUHK students, and a few pair examples from it are shown in Fig. 4. In our experiments, 88 sketch-photo pairs are randomly selected for training the model, and the remaining 100 pairs are used for testing. The recognition task is to identify the face photo corresponding to given a sketch. Similarly to the experiments on the Wikipedia dataset, the dictionaries learned during training are used to calculate the codes associated to test samples. Then, recognition is based on the computed codes.

Baselines. We compare the MPDL method to seven state-of-the-art methods, namely: CCA, Sketch Transform Method (STM) [31], PLS [25], [32], Bilinear model (Bil) [33], Semi-Coupled Dictionary Learning (SCDL) [2], Joint Dictionary Learning (JDL) [1] and Coupled Dictionary Learning (CDL) [3]. Similarly to MPDL, the SCDL, JDL and CDL methods are based on dictionary learning, and the dictionary size is set to 50 in all cases. For the bilinear model, we used 70 PLS bases and 50 eigenvectors (see [32]).

Results. Our model learns sparse representation from the raw sketch and face images. The regularization parameters α, β and γ were set by cross-validation to 1, 0.1 and 4.5 respectively. The dictionaries were initialized using joint dictionary learning [1]. The recognition is done using nearest neighbor classifiers, as in [32], [3], on the newly learned sparse representation.

Table II reports average recognition results over five trials. MPDL achieves the best average recognition rate: 98.4%. Remarkably, MPDL outperforms CDL, which is the best of the



(a) $\mathbb{Q}_{T\to I}$ (b) $\mathbb{Q}_{I\to T}$

Fig. 5. MAP for the cross-modal retrieval tasks as a function of the dictionary size for MPDL and MPDL ($\mu \to \infty$) on the Wikipedia dataset.

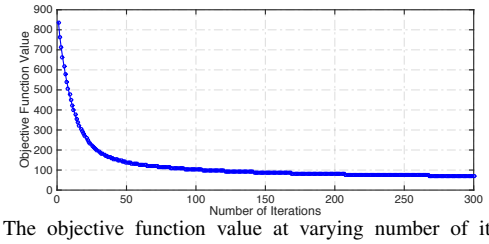


Fig. 6. The objective function value at varying number of iterations on Wikipedia dataset.

dictionary based approaches. Among the compared methods, SCDL, JDL and CDL are the strongest competitors, achieving 95.2%, 95.4% and 97.4% recognition rate respectively. This means that dictionary learning for cross-modal learning is a promising line of research further improved by the proposed MPDL. Specially, when the learning pace is driven by the data as shown by the 1.1% improvement over MPDL ($\mu \to \infty$).

C. Analysis of MPDL

In order to further analyze the proposed approach we first assess the performance as a function of the dictionary size. Figure 5 show MAP measures for the $\mathbb{Q}_{T\to I}$ and $\mathbb{Q}_{I\to T}$ tasks of MPDL and MPDL ($\mu \to \infty$) as a function of the dictionary size. Generally speaking, MPDL outperforms its non-paced version for most values of the dictionary size.

We also discuss and provide experimental evidence of the convergence of the learning algorithm. As illustrated in Section III-C, our MPDL relies on an alternating optimization approach and the proposed optimization problem (7) is solved separately with respect to dictionaries $\mathbf{D}^i$, codes $\mathbf{S}^i$ and self-paced variables $\mathbf{V}$. According to [34] the alternative search implemented in Algorithm 1 converges. Moreover, as illustrated in Fig. 6 for the Wikipedia dataset, we experimentally observe that MPDL attains a stable solution within few iterations, proving the efficiency of the algorithm proposed to solve the MPDL optimization problem. Regarding computational complexity, it is worth noting that for MPDL, as well as for any algorithm incorporating a SPL scheme, the improvement in terms of generalization performance comes at the price of an increased computational cost. However, in case of the considered retrieval and recognition tasks, the training time is not particularly important, as training is performed only once in an offline phase. Oppositely, the test phase, where codes are computed by projecting test samples into the learned dictionaries, is very efficient.

To conclude this in-depth analysis of the proposed method, we remark that both applications clearly show the advantage of using the cross-modal coupling term. In other words MPDL outperforms MPDL ($\gamma = 0$) by at least 2 points and up to 4 points (see Tables I and II).

V. Conclusions

This study introduces the multi-paced learning framework and an instantiation of it under dictionary learning. The rationale between multi-paced learning is that, when data from multiple modalities are available, the learning pace of each sample should be determined, not only by the reconstruction error (loss) in each modality but also from the coherence of the representations across modalities. We formulated the multi-paced dictionary learning problem and propose an efficient algorithm to solve it. An extensive evaluation campaign is performed on two cross-modal data analysis tasks, namely cross-modal retrieval and sketch-to-photo face recognition. Future works will focus on improving the proposed framework, *e.g.* exploiting an hybrid curriculum/multi-paced learning scheme [7], and on extending our multi-domain approach in case of other latent variables models.

References

[1] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, vol. 19, no. 11, pp. 2861–2873, 2010.

[2] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *CVPR*, 2012.

[3] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *ICCV*, 2013.

[4] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *NIPS*, 2010.

[5] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *ACM MM*, 2014.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009.

[7] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *AAAI*, 2015.

[8] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE T. Sig. Proc.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *ICML*, 2009.

[10] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *NIPS*, 2006.

[11] G. Liu, Y. Yan, E. Ricci, Y. Yang, Y. Han, S. Winkler, and N. Sebe, "Inferring painting style with multi-task dictionary learning," in *IJCAI*, 2015.

[12] Z. Ding, S. Ming, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *AAAI*, 2014.

[13] J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," in *CVPR*, 2013.

[14] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann, "Self-paced learning for matrix factorization," in *AAAI*, 2015.

[15] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *NIPS*, 2012.

[16] Y. Tang, Y.-B. Yang, and Y. Gao, "Self-paced dictionary learning for image classification," in *ACM MM*, 2012.

[17] C. Xu, D. Tao, and C. Xu, "Multi-view self-paced learning for clustering," in *IJCAI*, 2015.

[18] M. Zheng, J. Bu, C. Chen, C. Wang, L. Zhang, G. Qiu, and D. Cai, "Graph regularized sparse coding for image representation," *IEEE TIP*, vol. 20, no. 5, pp. 1327–1336, 2011.

[19] J. Song, Y. Yang, Z. Huang, H. T. Shen, and R. Hong, "Multiple feature hashing for real-time large scale near-duplicate video retrieval," in *ACM MM*, 2011.

[20] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *NIPS*, 2014.

[21] C. J. Albers, F. Critchley, and J. C. Gower, "Quadratic minimisation problems in statistics," *J. Mult. Anal.*, vol. 102, no. 3, pp. 698–713, 2011.
[22] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
[23] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM MM*, 2010.
[24] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *ACM MM*, 2003.
[25] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, latent structure and feature selection*, 2006, pp. 34–51.
[26] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *ACM MM*, 2013.
[27] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *IEEE CVPR*, 2010.
[28] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI*, 2011.
[29] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *PVLDB*, vol. 7, no. 8, pp. 649–660, 2014.
[30] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE TPAMI*, vol. 31, no. 11, pp. 1955–1967, 2009.
[31] X. Tang and X. Wang, "Face sketch recognition," *IEEE T. Circ. Syst. Video Tech.*, vol. 14, no. 1, pp. 50–57, 2004.
[32] A. Sharma and D. W. Jacobs, "Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch," in *IEEE CVPR*, 2011.
[33] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.
[34] D. Meng and Q. Zhao, "What objective does self-paced learning indeed optimize?" *arXiv preprint arXiv:1511.06049*, 2015.